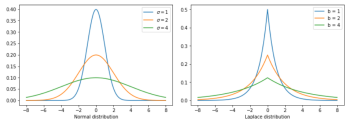


Introduction to Machine Learning

Regularization Bayesian Priors



Learning goals

- Know how regularized risk minimization is same as MAP in Bayesian perspective
- Know correspondence of Gaussian/Laplace priors and $L2/L1$ regularization

RRM VS. BAYES

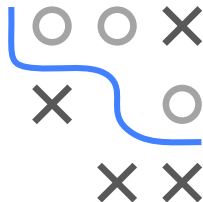
We already created a link between max. likelihood estimation and ERM.

Now we will generalize this for RRM.

Assume we have a parameterized distribution $p(y|\theta, \mathbf{x})$ for our data and a prior $q(\theta)$ over our parameter space, all in the Bayesian framework.

From the Bayes theorem we know:

$$p(\theta|\mathbf{x}, y) = \frac{p(y|\theta, \mathbf{x})q(\theta)}{p(y|\mathbf{x})} \propto p(y|\theta, \mathbf{x})q(\theta)$$



RRM VS. BAYES / 2

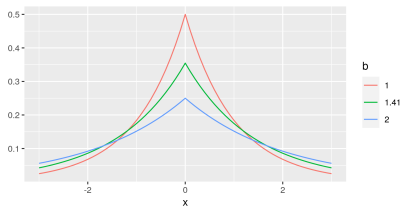
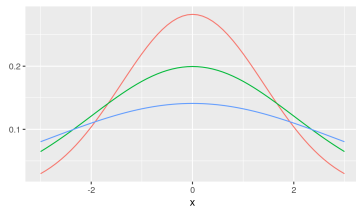
The maximum a posteriori (MAP) estimator of θ is now the minimizer of

$$-\log p(y | \theta, \mathbf{x}) - \log q(\theta).$$

- Again, we identify the loss $L(y, f(\mathbf{x} | \theta))$ with $-\log(p(y|\theta, \mathbf{x}))$.
- If $q(\theta)$ is constant (i.e., we used a uniform, non-informative prior), the second term is irrelevant and we arrive at ERM.
- If not, we can identify $J(\theta) \propto -\log(q(\theta))$, i.e., the log-prior corresponds to the regularizer, and the additional λ , which controls the strength of our penalty, usually influences the peakedness / inverse variance / strength of our prior.



RRM VS. BAYES / 3



- $L2$ regularization corresponds to a zero-mean Gaussian prior with constant variance on our parameters: $\theta_j \sim \mathcal{N}(0, \tau^2)$
- $L1$ corresponds to a zero-mean Laplace prior: $\theta_j \sim \text{Laplace}(0, b)$. $\text{Laplace}(\mu, b)$ has density $\frac{1}{2b} \exp(-\frac{|\mu-x|}{b})$, with scale parameter b , mean μ and variance $2b^2$.
- In both cases, regularization strength increases as the variance of the prior decreases: a prior probability mass more narrowly concentrated around 0 encourages shrinkage.
- Elastic-net regularization corresponds to a compromise between Gaussian and Laplacian priors ► Zou and Hastie 2005 ► Hans 2011

EXAMPLE: BAYESIAN L2 REGULARIZATION

We can easily see the equivalence of $L2$ regularization and a Gaussian prior:

- We define a Gaussian prior with uncorrelated components for θ :

$$q(\theta) = \mathcal{N}_d(\mathbf{0}, \text{diag}(\tau^2)) = \prod_{j=1}^d \mathcal{N}(0, \tau^2) = (2\pi\tau^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2\tau^2} \sum_{j=1}^d \theta_j^2\right).$$

- With this, the MAP estimator becomes

$$\begin{aligned}\hat{\theta}^{\text{MAP}} &= \arg \min_{\theta} (-\log p(y | \theta, \mathbf{x}) - \log q(\theta)) \\ &= \arg \min_{\theta} \left(-\log p(y | \theta, \mathbf{x}) + \frac{d}{2} \log(2\pi\tau^2) + \frac{1}{2\tau^2} \sum_{j=1}^d \theta_j^2 \right) \\ &= \arg \min_{\theta} \left(-\log p(y | \theta, \mathbf{x}) + \frac{1}{2\tau^2} \|\theta\|_2^2 \right).\end{aligned}$$

- We see how the inverse variance (precision) $1/\tau^2$ controls shrinkage.



EXAMPLE: BAYESIAN L2 REGULARIZATION / 2

- Consider a DGP $y = \theta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, 1)$ and $\theta = 1$ together with a Gaussian zero-mean prior on θ , i.e., $\mathcal{N}(0, \tau^2)$ for $\tau \in \{0.25, 0.5, 2\}$
- Given $n = 20$ samples, the posterior of θ and its MAP estimate can be calculated analytically
- Plotting the L_2 regularized empirical risk $\mathcal{R}_{\text{reg}}(\theta) = \sum_{i=1}^n (y_i - \theta)^2 + \lambda \theta^2$ with $\lambda = 1/\tau^2$ shows that ridge solution is identical with MAP